The paper presents an application of deep learning techniques, specifically U-Net and GAN-based models, to enhance short-term precipitation forecasting, with a focus on the fine-scale structure of intense rainfall events. The authors compare the accuracy of three numerical weather prediction (NWP) models against two deep learning techniques: a U-Net-based model (FRNet) and a GAN-based model (GFRNet). They use multiple evaluation metrics to assess the relative performances of these approaches. The goal is to evaluate the accuracy of predicting 3-hourly accumulated precipitation over the next 24 hours for a region in North China. The paper shows many metrics and concludes that GFRNet demonstrates significant operational value.

0.1 Question 1

The paper deals with an interesting topic, related to improving the forecasting of the fine structure of intense rainfall. Moreover, it makes use of some current tools in deep learning, which seem promising for future operational use. However, I have two main concerns about the paper. The first one is the rationale of the model itself, as I do not understand how the fine structure of rainfall may be solely explained with the additional information (on top of the NWPs) provided to the GAN. The second one is related to the experimental design and the fairness of the comparisons and analyses provided. In my opinion, the paper requires an improvement of the rationale and the experimental design, as well as additional analyses before being considered for publication in a scientific journal.

0.2 Answer 1

Thank you for your thoughtful comments and constructive feedback on our paper. Regarding your concern about how the GAN-based model improves the fine structure of precipitation forecasts:

1. As introduced in the Introduction and Method sections, the principle of GAN lies in the adversarial training between the generator and discriminator. The goal is to train the generator to produce outputs that are indistinguishable from real labels, as judged by the discriminator. This process effectively allows the generator to refine the predictions to a level that achieves "realistic" quality.

2. In the field of short-term precipitation forecasting, the scientific validity of GANs has been demonstrated in two prominent works published in Nature (Ravuri et al., 2021; Zhang et al., 2023). These studies highlight GANs' ability to capture complex spatial and temporal structures, making them a promising tool for fine-scale rainfall prediction. Building on these findings, our work explores the application of GANs specifically for short-term precipitation forecasting.

I hope this addresses your concern. Please let me know if you would like further clarification or additional details.

0.3 Question 2

Below, I provide more details about the issues that I observe with the work.

My main concern with the rationale of the paper is that it is not immediately evident how the information about Elevation, Latitude, Longitude, Cycle, and Lead Hour contributes independently to improving forecasts when much of this information may already be embedded in the NWPs. A mechanism should be presented or outlined to justify the gains in accuracy. If none exists, then all the required information about the fine structure of precipitation is already included in the original NWPs, and thus the methods presented are just extracting this information.

0.4 Answer 2

Thank you for your insightful comments regarding the rationale of our paper. Below, we address your concerns regarding the contributions of elevation, latitude, longitude, cycle, and lead hour information to improving forecasts:

1. While it is true that NWPs already encapsulate fine-scale precipitation structures, each individual NWP model exhibits specific strengths and weaknesses. The goal of our deep learning model is to dynamically integrate these complementary features in a way that achieves a synergistic effect, effectively making 1 + 1 > 1.

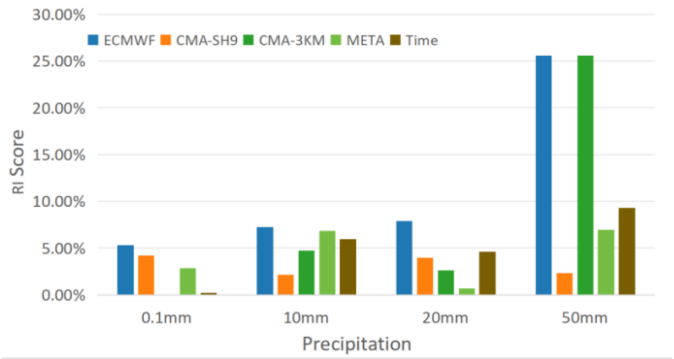

Figure 1. The Relative Importance Score of the input features.

2. Regarding the contributions of geographical (elevation) and temporal (cycle, lead hour) encodings relative to other NWP features, we conducted ablation studies in a prior work (currently under review) to quantify the importance of these features. Using FRNet as the baseline model (note: due to differences in test sets, specific values may vary slightly. See Table 1), we defined the Relative Importance Score (RIScore) as follows Equation 1 .The results demonstrated that while the precipitation outputs from NWPs contribute the most to forecast improvements, geographical and temporal features also play a meaningful role (See Figure 1).

$$\text{Relative Importance}(x) = \frac{\text{TS(FRNet)} - \text{TS(wo\_x)}}{\text{TS(FRNet)}} \tag{1}$$

3. These findings illustrate that deep learning models not only effectively integrate precipitation forecasts from multiple NWPs but also have the capacity to learn hidden mapping relationships from features closely tied to precipitation, such as terrain and time, thereby further enhancing forecast accuracy.

Table 1. TS and BIAS scores for different rain thresholds in ablation experiments

Model/Threshold	TS_0.1	TS_10	TS_20	TS_50	BIAS_0.1	BIAS_10	BIAS_20	BIAS_50
wo_ECMWF	0.429	0.217	0.14	0.032	1.094	1.655	1.827	0.931
wo_CMA-SH9	0.434	0.229	0.146	0.042	0.858	1.598	1.823	1.973
wo_CMA-3KM	0.453	0.223	0.148	0.032	1.203	1.708	1.838	1.229
wo_META	0.44	0.218	0.151	0.04	**0.995**	1.545	**1.36**	0.689
wo_Time	0.452	0.22	0.145	0.039	1.078	1.504	1.706	**1.065**
FRNet	**0.453**	**0.234**	**0.152**	**0.043**	1.063	**1.476**	**1.392**	0.781

We hope this explanation provides clarity. Please let us know if additional details or analyses are required.

0.5 Question 3

If this is the case, as I believe (although I may be wrong), there may be other alternative methods that could improve forecasting with reduced complexity. To verify this point, I suggest that the authors include additional models, such as SVMs or Random Forests, to test if simpler bagging methods with far fewer parameters could also improve forecasting accuracy. In my experience, basic machine learning methods tend to perform similarly to deep learning ones (in this kind of application) at *a significantly reduced level of complexity. Including these additional comparisons may serve to justify whether a GAN is an optimal strategy and to show if the improvement in forecasting accuracy comes from the deep learning techniques or from the combination of different sources of information.*

Using FRNet as a benchmark may not be entirely fair, as GFRNet is essentially an enhanced version of the same model with a more advanced training procedure. It would also be interesting to report training and inference times for all models used (I know that the authors have included part of this information in their manuscript).

Thus, with respect to my concern with the rationale of the paper, the authors should include additional simpler models to check if GANs are justified for their complexity level or if other simpler methods may work similarly. Additionally, they should provide some insight into why the improvement occurs. This point leads to my concerns with the experimental design.

0.6 Answer 3

Thank you for your detailed feedback and suggestions regarding the experimental design and the rationale for using GANs in our study. Below, we address your concerns point by point:

1. **Comparison with Simpler Models**. Numerous existing studies have applied machine learning models, such as SVMs or Random Forests, to precipitation correction tasks. These methods, with careful design and parameter tuning, have been shown to outperform NWPs in accuracy. We acknowledge that simpler models with reduced complexity and computational efficiency could achieve high accuracy. However, a common limitation across these non-generative models is the blurring effect in their predictions, as highlighted in prior works.

2. **Rationale for Using GANs**.The primary motivation of our work is to explore whether adopting a generative approach, such as GANs, can address this blurring issue without compromising accuracy. By introducing adversarial training, we aim to produce sharper and more realistic precipitation predictions, especially for fine-scale structures.

3. **Fair Comparison Between Paradigms**.To ensure a fair comparison between non-generative and generative paradigms, we used the same model architecture (FRNet) for the generator component of the GAN. This approach isolates the impact of the adversarial training process, providing a controlled evaluation of the benefits of generative modeling over traditional methods.

4. **Combining Information Sources.** While the quality and diversity of data sources play a critical role, effectively combining these inputs requires leveraging advanced AI techniques. Our ablation studies have demonstrated that the deep learning model in our study successfully integrates these features to achieve positive correction results. It is important to note that the performance of these post-processing models inherently depends on the quality of NWPs. Thus, advancements in both NWPs and deep learning techniques can collectively enhance post-processing outcomes.

We appreciate your suggestion to include simpler models for comparison. While this would undoubtedly add value, the primary focus of our study is to address the limitations of non-generative models and evaluate the advantages of GAN-based techniques. Please let us know if additional analyses or clarifications are required.

0.7 Question 4

My first concern with the experimental design is related to the limited amount of data used in the study. Using only four years of data may not be sufficient to fully verify the accuracy and robustness of the forecasting method. I understand that data limitations are difficult to overcome, but given the complexity of the models used, it is difficult to ensure that overfitting is not playing a role in the analysis. This concern is exacerbated by the fact that the original time series must be split into training, validation, and test sets.

Moreover, the initial data selection may bias the results. NWPs provide continuous forecasts, so alternative methods should ideally also deliver continuous predictions to ensure a fair comparison. If a heavy data selection procedure is implemented, the comparison may not be entirely fair to the NWPs.

0.8 Answer 4

Thank you for raising this important concern regarding the amount of data and the potential for overfitting in our study. We acknowledge the limitations inherent in using four years of data and have taken steps to ensure the robustness and fairness of our experimental design. Below, we address your points in detail:

1. **Data Limitations and Generalization**.We agree that collecting and preparing sufficient data for deep learning models is challenging and resource-intensive. It is indeed difficult to guarantee that the model will maintain strong generalization capability on entirely new datasets, as patterns in new data may change. This is a common challenge faced by all AI methods.

2. **Principles for Dataset Partitioning**.Given the constraints of limited data, it is essential to follow strict principles to ensure that the model's performance is both accurate and robust: a.**There should be no data leakage** between the training, validation, and test sets. b.**The test set should be unbiased**, representing real-world scenarios as comprehensively as possible.

3. **Our Dataset Partitioning and Testing Strategy**. **a. No Data Leakage:** As shown in Table 2, our dataset was partitioned chronologically, ensuring no overlap between training, validation, and test sets, thereby eliminating the risk of data leakage. **b. Unbiased and Realistic Test Set**: While we employed a sampling strategy on the training set to address data imbalance, the validation and test sets consisted of continuous, unaltered natural samples. Specifically, the test set included 77 consecutive days of real-world summer weather samples, providing an unbiased representation of the study region's conditions. **c. Validation of Generalization**: The stopping criterion for model training was defined by the minimum loss observed on the validation set, with no further improvement for 20 consecutive epochs. Good performance on the validation set is, therefore, not unexpected. However, if the model performs well on the validation set but poorly on the test set, it would indicate overfitting and poor generalization.

In our study, all evaluation results were conducted on the independent test set, and the model demonstrated strong performance. This suggests that, within the limitations of the available data, the model possesses good generalization capability on unseen data.

Table 2. Sample distribution across training, validation, and test sets.

Dataset	Time Period	Samples	
		Pre-sampling	Post-sampling
Training set	2019-06-01 - 2019-10-10 2020-06-01 - 2020-10-10 2021-03-15 - 2021-07-09 2021-08-21 - 2021-10-10 2022-03-15 - 2022-06-14	4645	2885
Validation set	2021-07-10 - 2021-08-20	637	No sampling
Test set	2022-06-16 - 2022-08-31	1204	No sampling

0.9 Question 5

A second concern with the experimental design is that I would have appreciated a clearer discussion in the methods' section about how the evaluation of accuracy was performed. Table 3 presents evaluation metrics and highlights the best and second-best performers. However, more attention should be paid to the differences. Are they significantly different? Or could all the methods (NWPs and nets) perform similarly given the amount of information used? What would be the expected distribution of *the accuracy metrics? I am not fully convinced that part of the results are not an analysis of statistical fluctuations. Additionally, Table 4 seems to contradict the abstract, which states that "GFRNet outperforms all models in terms of Root Mean Square Error (RMSE)," but I may have missed something.*

0.10 Answer 5

Thank you for your detailed review and insightful questions. Below, I address your concerns regarding the evaluation metrics, their significance, and the apparent discrepancy in Table 4.

Evaluation Metrics and Fairness. For short-term hourly precipitation forecasting, metrics such as TS, FAR, BIAS, and FSS are widely recognized and commonly used evaluation standards. In this study, these metrics were computed using standard statistical formulas and were evaluated under the same spatial and temporal resolutions to ensure fairness across all methods.

Significance of Differences. To determine whether the observed differences are meaningful, two key questions need to be addressed:

1. Are the differences statistically significant?
2. If so, how can we confirm that these differences are not due to statistical fluctuations but instead reflect the true performance of the models?

Model Performance Gains. Considering recent advancements in NWPs, achieving improvements in hourly short-term forecasts for heavy or torrential rainfall is exceptionally challenging. The TS and FSS improvements achieved by NETs (deep learning models) over the best-performing NWPs were approximately 20% for moderate, heavy, and torrential rainfall. Such improvements are statistically significant and represent meaningful performance gains.

Temporal and Spatial Analyses. We also analyzed the differences between NETs and NWPs in more detail from the two dimensions of time and space.

1. **Temporal Analysis:** As shown in Figure 4, we analyzed the TS and FSS scores of NWPs and NETs across different forecast lead times. The results demonstrate stable performance for all five models, particularly for thresholds of 10 mm and 20 mm. GFRNet and FRNet consistently showed significant advantages over NWPs.
2. **Spatial Analysis:** Figures 5 and 6 illustrate that GFRNet outperformed NWPs in most forecast regions, highlighting its spatial robustness. Additionally, since NETs are post-processed from NWPs, the performance trends of NETs and NWPs over forecast lead times exhibit similar patterns.

Clarification on Table 4 and Abstract.Thank you for pointing out the discrepancy regarding RMSE. You are correct that GFRNet does not achieve the lowest RMSE; ECMWF has the lowest RMSE. However, ECMWF's low RMSE is primarily due to its conservative predictions for moderate and heavy rainfall, which lack forecasting skill. In contrast, GFRNet achieves the second-lowest RMSE while maintaining high forecast skill, demonstrating a balance between accuracy and robustness. We will revise the abstract to accurately reflect this point.

0.11 Question 6

A third concern is related to the case studies, which I believe should be justified and presented in a more detailed way. This point may be related to my concern about the rationale of the paper: if a mechanism by which the GAN strategy improves the forecast is provided, then the case studies may focus on clear examples of this mechanism at play. Without this, I believe a general statistical evaluation would provide a clearer representation of the model's advantages. A detailed analysis of specific situations may not be as illuminating.

0.12 Answer 6

Thank you for your thoughtful comments regarding the case studies. We acknowledge the importance of a well-justified and detailed presentation of case studies, particularly in demonstrating the mechanisms by which the GAN strategy improves forecasts.

While general statistical evaluations provide an overall representation of model performance, they may obscure certain limitations or strengths of the models. Based on suggestions from meteorological experts, we selected three distinct precipitation events characterized by different dynamic and thermodynamic conditions. By visualizing these cases and providing detailed case-specific scores, we aimed to illustrate the stability of NETs, particularly GFRNet, across various precipitation scenarios.

Moreover, compared to FRNet, which tends to produce blurred predictions, GFRNet consistently delivers forecasts with clearer precipitation structures in these case studies. This highlights the GAN-based model's ability to address the blurring issue and capture fine-scale precipitation details more effectively.

We believe these case studies complement the statistical analysis by providing deeper insights into the model's performance under diverse conditions.

0.13 Question 7

A fourth concern about the experimental design is related to the selection of three NWPs. If two models similar to ECMWF were available, would it make sense to include both? How would the results change? This raises questions about generalizability. In many machine learning applications, the data exert a closer control on accuracy than the algorithms themselves. I understand that a paper cannot address every concern, but some guidance from the authors would be appreciated.

0.14 Answer 7

Thank you for raising this insightful question. The selection of NWPs and its impact on generalizability is indeed a critical consideration.

From the principles of ensemble forecasting design and prior studies on multi-model post-processing, an important guideline for selecting ensemble members is that they should be "**high-quality and diverse.**" This ensures that the input information provided to the correction model is both accurate and comprehensive, maximizing the potential for improved forecast performance. If two models similar to ECMWF were included, they would likely offer redundant information, providing little additional benefit beyond what a single ECMWF model could contribute. Consequently, the improvements in forecast accuracy might be limited.

Regarding the relationship between data and algorithms, we share your perspective that data defines the upper limit of performance, while the iterative improvement of models seeks to approach this limit. For smaller datasets, simpler models may suffice, whereas larger datasets often require more complex models to fully exploit the information available. The development of both data and models should ideally evolve in tandem to achieve optimal results.

We hope this addresses your concern. Please let us know if further clarification or additional discussion would be helpful.

0.15 Question 8

My final concern relates to the generality of the conclusions and the reproducibility of the results in other locations. How robust are the results to the data selection procedure or the structure of the ANN? Would the same structure work well in other locations, or would changes be required? If a less intense data selection procedure were used, how would the results change? If 20 years of data were available, would GFRNet perform similarly? I believe the study would be much more robust if extended to other regions with more data available. Currently, the method seems to work, but the evidence may not yet be robust enough to fully support the claims made in the paper.

0.16 Answer 8

Thank you for your thoughtful comments and for raising concerns about the generality and reproducibility of the study's conclusions. These are critical points that deserve careful discussion.

Currently, due to the substantial effort required for data collection and preprocessing, we have not yet validated the generalization capability of the model trained in the North China region to other geographic locations. Precipitation patterns are indeed highly region-specific, influenced by local geography, climate, and dynamics.

That said, one of the key contributions of our study lies in proposing and demonstrating a methodology: **using generative deep learning models to improve both the accuracy and fine-scale structure of precipitation forecasts**. This approach, while tailored to the North China region, is intended to serve as a guiding framework. With this methodology, similar models could be trained or fine-tuned for other regions or extended datasets, such as longer time periods or more extensive geographic areas. We are optimistic that this approach would yield comparable results, though further studies are needed to verify this.

If more extensive datasets, such as 20 years of data, were available, we believe GFRNet would continue to perform well. A larger dataset could enable the model to better capture long-term patterns and variability, potentially improving its robustness and generalizability further.

We appreciate your suggestions and agree that extending this work to other regions with more data would significantly strengthen the study's conclusions. Such an extension is a valuable direction for future research.

0.17 Question 9

Finally, I present some comments about minor issues:

1. *References should be enclosed in parentheses. The way they are written now complicates the reading of the paper.*
2. *Figures 4 and 5 are difficult to interpret, particularly due to their complex visual layout. A more intuitive representation could enhance their clarity. For the maps, since topography seems to play such an important role, residuals might provide better insights.*
3. *Some references to equations are incomplete.*
4. *A better discussion on ensemble forecasts and deterministic quantitative forecasts may be in order. In my opinion, ensemble forecasts may convey a much better idea of severe storm potential, especially when combined with synthetic generation, so focusing on deterministic forecasts may be a disadvantage.*

0.18 Answer 9

Thank you for your detailed comments and suggestions regarding minor issues. Below, we address each of your points:

1. **References.** We appreciate your observation regarding the formatting of references. We will revise the manuscript to ensure that all references are properly enclosed in parentheses, improving the consistency and readability of the text.
2. **Figures 4 and 5**. Thank you for your comments on Figures 4 and 5. To address their complexity, we will simplify the visual layout and explore more intuitive representations. Indeed, residual maps can better illustrate the impact of topography on precipitation forecasts. As such, we have included the spatial distribution of residuals under the FSS metric in Figure 6, comparing GFRNet with NWPs and FRNet, to provide deeper insights into this effect.
3. **Equations.** Thank you for pointing out the incomplete references to equations. We will carefully review the manuscript to ensure all equation references are complete and formatted correctly.
4. **Discussion on Ensemble Forecasts.** We agree that ensemble forecasts have significant advantages in capturing the potential for severe weather events, particularly when combined with synthetic generation techniques. However, the primary focus of this study is to evaluate the capability of deterministic forecasts in predicting precipitation structures and fine-scale details. In future work, we plan to explore how ensemble forecasting techniques can complement or enhance our deterministic approach, especially for severe storm scenarios.

We greatly appreciate your thoughtful feedback and will incorporate these suggestions to improve the manuscript. Please let us know if there are additional areas that require further attention.

References

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skillful Precipitation Nowcasting Using Deep Generative Models of Radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Zhang, Y., Long, M., Chen, K., Xing, L., Jin, R., Jordan, M. I., and Wang, J.: Skilful Nowcasting of Extreme Precipitation with NowcastNet, Nature, pp. 1–7, https://doi.org/10.1038/s41586-023-06184-4, 2023.